Thermal Guided Extreme Low Light Image Enhancement for ADAS Applications

Raman Jha [1] Adithya Lenka [1] Mani R. [2] A. C. Sankaranarayanan [2] K. Mitra [1]

[1]Indian Institute of Technology, Madras [2]Carnegie Mellon University, Pittsburgh

Introduction

Vehicle accidents, almost all of which are caused by human error, can be avoided by employing Advanced Driver Assistance Systems (ADAS). ADAS applications such as pedestrian detection and traffic sign recognition make use of deep learning models trained on computer vision algorithms. The use of such models during night time in particular is a challenge due to the severe degradation, such as contrast/color reduction and noise deterioration, that typically occurs as a result of insufficient ambient illumination. We propose that through the use of thermal imagery with its consistent information capture regardless of illumination and light variations, the enhancement of low-light visible image can be benefited.

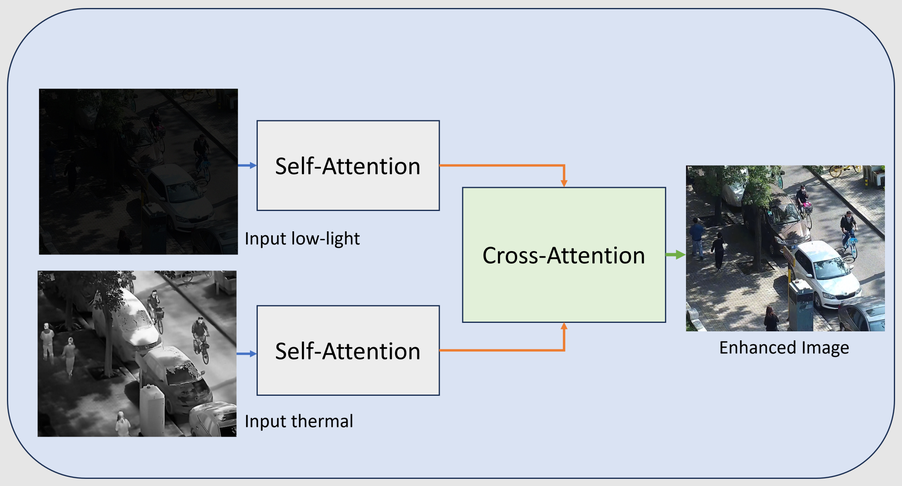
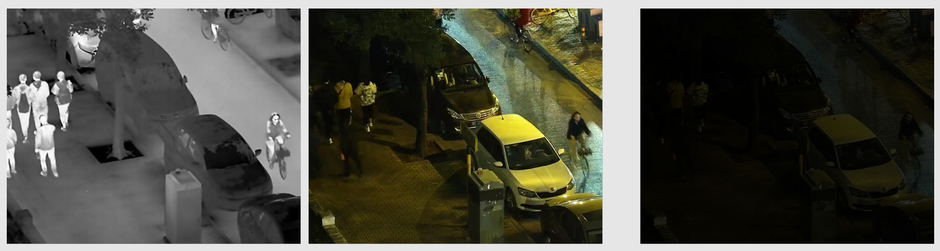


Figure 1. Overview

Data

We make use of the LLVIP dataset, a visible-infrared paired dataset and simulate low-light on the well-exposed night-time images. We train our proposed model and fine-tune pre-existing state-of-the-art visible low-light image enhancement models on the synthetic low-light LLVIP dataset. While simulating low-light by reducing image exposure by factors ranging from 5 to 20, we introduced noise using the method outlined by Hasinoff et al. 2010.



Thermal Image | Ground-truth | Simulated Low Light

Figure 2. Synthetic LLVIP dataset

We test the above models on a real-world dataset collected by collaborators from Carnegie Mellon University, Pittsburgh.

Methodology

- In Retinex theory, a low-light image I with dimensions H×W×3 can be decomposed into two constituent components: a reflectance image R with identical dimensions as I, and an illumination map M with dimensions H×W. Thus, the image I can be represented as:

$$I = R \odot M \tag{1}$$

  Here, $\odot$ represents the element-wise multiplication of the reflectance image and the illumination map.
  Taking inspiration from Retinexformer we make use of an **illumination module** consisting of convolutional layers that output the illumination map and illumination features of the input images.

- Since CNN-based methods, due to their local receptive fields, show limitations in capturing long-range dependencies of different regions, we employ a transformer-based architecture that employs **self attention** (Kolesnikov et al. 2021) on the features of the visible and thermal images separately.
  Mathematically, attention is given as:

$$Attention(Q, K, V) = Softmax(\frac{Q \cdot K^T}{\sqrt{d_k}}) \cdot V \tag{2}$$

- We employ **cross attention** (Chen et al. 2021) mechanism on the transformer architecture to enable the model to fuse the information from the thermal and RGB modalities. The thermal features which retain the structural information of the scene provide the query for the attention mechanism, while the key and value tokens are from the RGB input image.

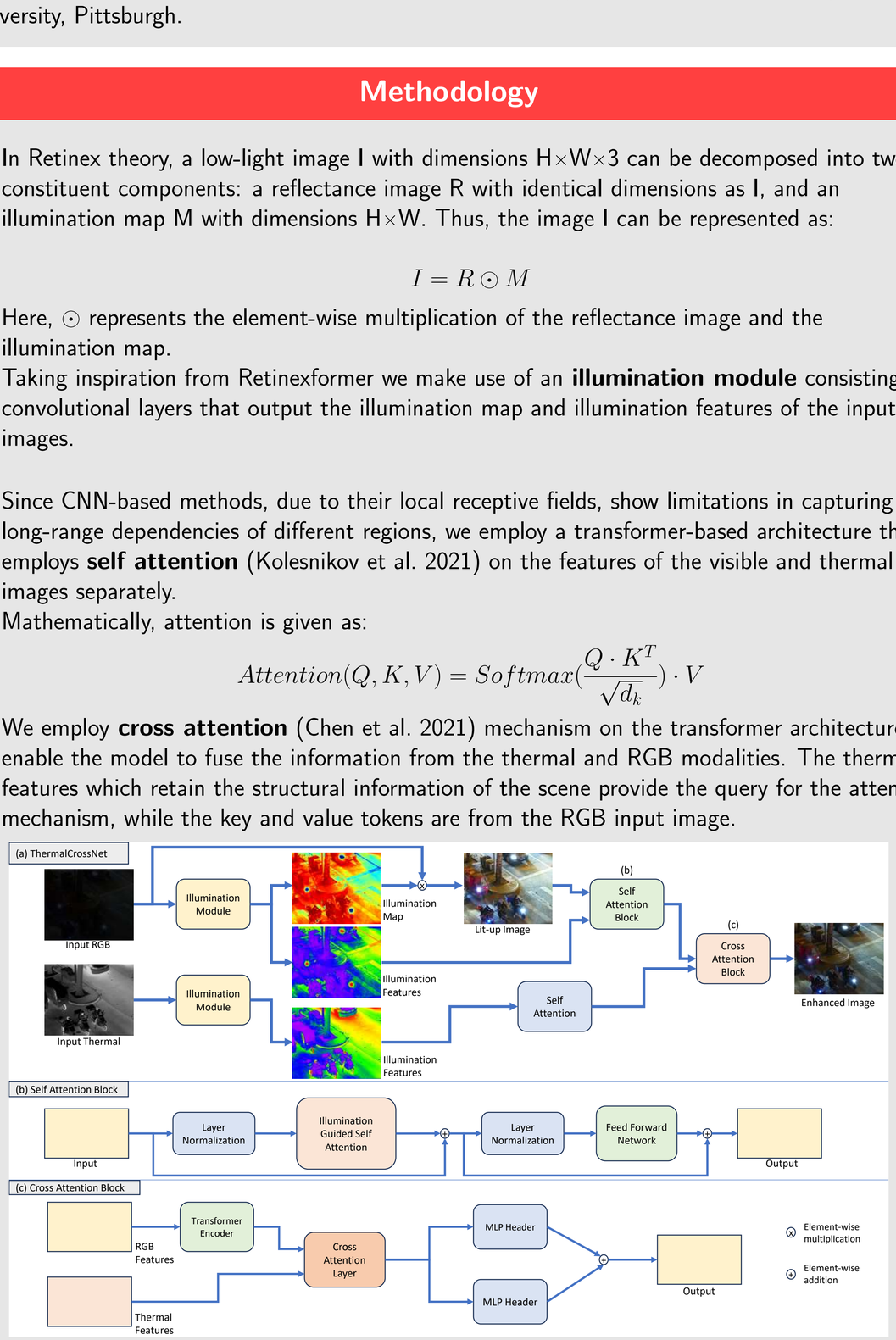


Figure 3. Model Architecture

Results

We compare the performance of the proposed model with state-of-the-art RGB only low-light image enhancement models. We adopt Peak Signal to Noise Ratio (PSNR) and Structure Similarity Index (SSIM) metrics to for the LLVIP test subset and Learned Perceptual Image Patch Similarity (LPIPS) and SSIM on our real-world collected dataset.

Method	LLVIP		real-world V-TIEE	
	PSNR ↑	SSIM ↑	LPIPS ↓	SSIM ↑
EnGAN	25.67	0.74	0.17	0.65
Retinexformer	26.59	0.79	0.14	0.66
ThermalCrossNet	**27.75**	**0.85**	**0.12**	**0.71**

Table 1. Quantitative comparison on LLVIP and real-world dataset

We also conduct a qualitative assessment of the output of our proposed model with the ground truth image for LLVIP dataset and with the well exposed reference image for our real-world collected dataset.

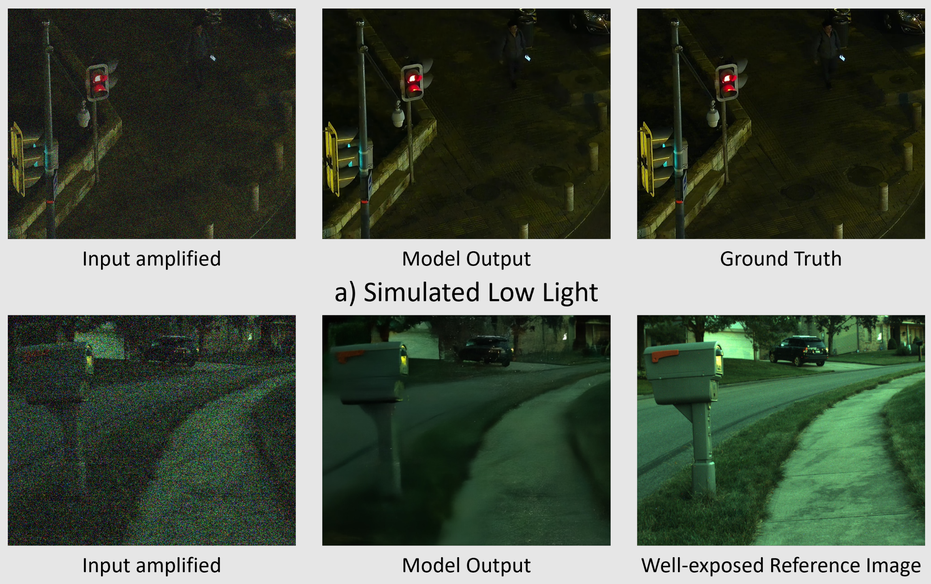


Figure 4. Qualitative Assessment

We further conduct an ablation study to evaluate the advantage in performance provided by the thermal image and our novel cross attention block.

Ablation	PSNR ↑	SSIM ↑
Self-Attention(Only RGB)	26.42	0.73
Thermal channel concatenation	27.15	0.79
Cross-Attention	**28.57**	**0.85**

Table 2. Ablation Study on LLVIP dataset

Conclusion

- Our model is able to make use of the information from the thermal images and outperform RGB-only state-of-the-art Low-Light Image Enhancement models.
- It is able to generalize to our real-world collected dataset without any fine-tuning and outperforms other models trained on the same data.

References

Chen, C., Q. Fan, and R. Panda (Oct. 2021). "CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification". In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 347–356. DOI: 10.1109/ICCV48922.2021.00041.

Kolesnikov, A., A. Dosovitskiy, D. Weissenborn, G. Heigold, J. Uszkoreit, L. Beyer, M. Minderer, M. Dehghani, N. Houlsby, S. Gelly, T. Unterthiner, and X. Zhai (2021). "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In.

Hasinoff, S. W., F. Durand, and W. T. Freeman (2010). "Noise-optimal capture for high dynamic range photography". In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 553–560. DOI: 10.1109/CVPR.2010.5540167.